

CUNY SPS DATA 621 - CTG5 - Final

Betsy Rosalen, Gabrielle Bartomeo, Jeremy O'Brien, Lidiia Tronina, Rose Koh

May 23rd, 2019

Contents

1	PROJECT DESCRIPTION AND BACKGROUND	2
1.1	Background	2
1.2	Hypothesis	2
1.3	Our approach, setup, and workflow	2
2	DATA PREPARATION	3
2.1	Description	3
2.2	Cross-validation	3
2.3	Bootstrap surrogate data	3
2.4	Synthesis diagnostics	3
3	BUILDING MODELS	4
3.1	Logistic regression	4
3.2	Decision tree (CHAID or C&RT?)	4
3.3	Random forest	4
3.4	Support Vector Machines	4
3.5	Naive Bayes	4
4	MODEL REVIEW AND SELECTION	5
4.1	Comparison of performance between models	5
4.2	Comparison of performance viz. other studies	5
5	CONCLUSIONS	6
6	APPENDIX	7
6.1	Supplemental tables and figures	7
6.2	R statistical programming code	13

1 PROJECT DESCRIPTION AND BACKGROUND

1.1 Background

With nearly 18MM deaths in 2015 , cardiovascular diseases (CVD) are the leading cause of death globally and growing in the developing world . CVD is a disease class which includes heart attacks, strokes, heart failure, coronary artery disease, arrhythmia, venous thrombosis, and other conditions. About half of all Americans (47%) have at least one of three key risk factors for heart disease: high blood pressure, high cholesterol, and smoking.

Researchers estimate that up to 90% of heart disease deaths could be prevented. Typical means of detection include electrocardiograms (ECGs), stress tests, and cardiac angiograms, all of which are expensive. Risk evaluation screenings require blood samples, which are assessed alongside risk factors like tobacco use, diet, sleep disorders, physical inactivity, air pollution, and others.

More efficient, scalable, and non-invasive means of early detection could be used to trigger medical interventions, prompt preventive care by physicians, and/or engender behavioral change on the part of those prone to or suffering from CVD. Applying data mining techniques to CVD datasets to predict risk based on existing or easy-to-collect health data could improve healthcare outcomes and mortality rates.

The Cleveland dataset is the most complete CVD dataset and is the most frequently used in data science experimentation. It has a small number of observations, but this is not uncommon given the costs of experimental data and privacy risk of observational data.

1.2 Hypothesis

1.3 Our approach, setup, and workflow

[We should discuss this]

2 DATA PREPARATION

2.1 Description

[Jeremy's adding in the variable description table from our proposal doc]

2.2 Cross-validation

[Assuming this requires a little explanation]

2.3 Bootstrap surrogata data

[Jeremy's writing this up]

Per the synthpop packace explanation (<https://cran.r-project.org/web/packages/synthpop/vignettes/synthpop.pdf>): “The basic idea of synthetic data is to replace some or all of the observed values by sampling from appropriate probability distributions so that the essential statistical features of the original data are preserved. The approach has been developed along similar lines to recent practical experience with multiple imputation methods although synthesis is not the same as imputation. Imputation replaces data which are missing with modelled values and adjusts the inference for the additional uncertainty due to this process. For synthesis, in the circumstances when some data are missing two approaches are possible, one being to impute missing values prior to synthesis and the other to synthesise the observed patterns of missing data without estimating the missing values. In both cases all data to be synthesised are treated as known and they are used to create the synthetic data which are then used for inference. The data collection agency generates multiple synthetic data sets and inferences are obtained by combining the results of models fitted to each of them. The formulae for the variance of estimates from synthetic data are different from those used for imputed data.”

“Our aim in writing the synthpop package (Nowok, Raab, Snoke, and Dibben 2016) for R (R Core Team 2016) is a more modest one of providing test data for users of confidential datasets. Note that currently all values of variables chosen for synthesis are replaced but this will be relaxed in future versions of the package. These test data should resemble the actual data as closely as possible, but would never be used in any final analyses. The users carry out exploratory analyses and test models on the synthetic data, but they, or perhaps staff of the data collection agencies, would use the code developed on the synthetic data to run their final analyses on the original data. This approach recognises the limitations of synthetic data produced by these methods.”

2.4 Synthesis diagnostics

The original Cleveland dataset contains $n = 303$ observations over ...

3 BUILDING MODELS

Our literature review revealed that of the many approaches that have been taken, certain types of models stand out in terms of their performance.

[Jeremy's including grid with examples of previous work]

[See notes in literature reviewed and Kaggle projects for suggestions on variable selection and feature engineering for models (links TBC)]

[See also notes on setup and packages for decision tree, random forest, SVM, and Naive Bayes (links TBC)]

3.1 Logistic regression

3.2 Decision tree (CHAID or C&RT?)

3.3 Random forest

3.4 Support Vector Machines

3.5 Naive Bayes

4 MODEL REVIEW AND SELECTION

4.1 Comparison of performance between models

[sensitivity, specificity, accuracy, others metrics?]

[Between different techniques and between original dataset and synthesized dataset for each technique]

4.2 Comparison of performance viz. other studies

5 CONCLUSIONS

6 APPENDIX

6.1 Supplemental tables and figures

Table 1: Summary statistics for numerical variables in the original dataset

	n	min	mean	median	max	sd
age	303	29	54.366337	55.0	77.0	9.082101
trestbps	303	94	131.623762	130.0	200.0	17.538143
chol	303	126	246.264026	240.0	564.0	51.830751
thalach	303	71	149.646865	153.0	202.0	22.905161
oldpeak	303	0	1.039604	0.8	6.2	1.161075

Table 2: Summary statistics for categorical variables in the original dataset

	cp	ca	restecg	slope	thal
	0:143	0:175	0:147	0: 21	0: 2
	1: 50	1: 65	1:152	1:140	1: 18
	2: 87	2: 38	2: 4	2:142	2:166
	3: 23	3: 20	NA	NA	3:117
	NA	4: 5	NA	NA	NA

Table 3: Summary statistics for binary categorical variables in the original dataset

	exang	fbs	sex	target
	0:204	0:258	0: 96	0:138
	1: 99	1: 45	1:207	1:165

Table 4: Summary statistics for numerical variables in the synthesised dataset

	n	min	mean	median	max	sd
age	6060	29	54.345875	55.0	77.0	9.103874
trestbps	6060	94	131.863201	130.0	200.0	17.401938
chol	6060	126	246.852805	242.0	564.0	52.916827
thalach	6060	71	149.968317	153.0	202.0	23.242878
oldpeak	6060	0	1.041535	0.8	6.2	1.165680

Table 5: Summary statistics for categorical variables in the synthesised dataset

	cp	ca	restecg	slope	thal
	0:2846	0:3448	0:3041	0: 504	0: 46
	1:1007	1:1363	1:2955	1:2783	1: 396
	2:1740	2: 764	2: 64	2:2773	2:3339
	3: 467	3: 370	NA	NA	3:2279
	NA	4: 115	NA	NA	NA

Table 6: Summary statistics for binary categorical variables in the synthesised dataset

	exang	fbs	sex	target
	0:4099	0:5099	0:1860	0:2875
	1:1961	1: 961	1:4200	1:3185

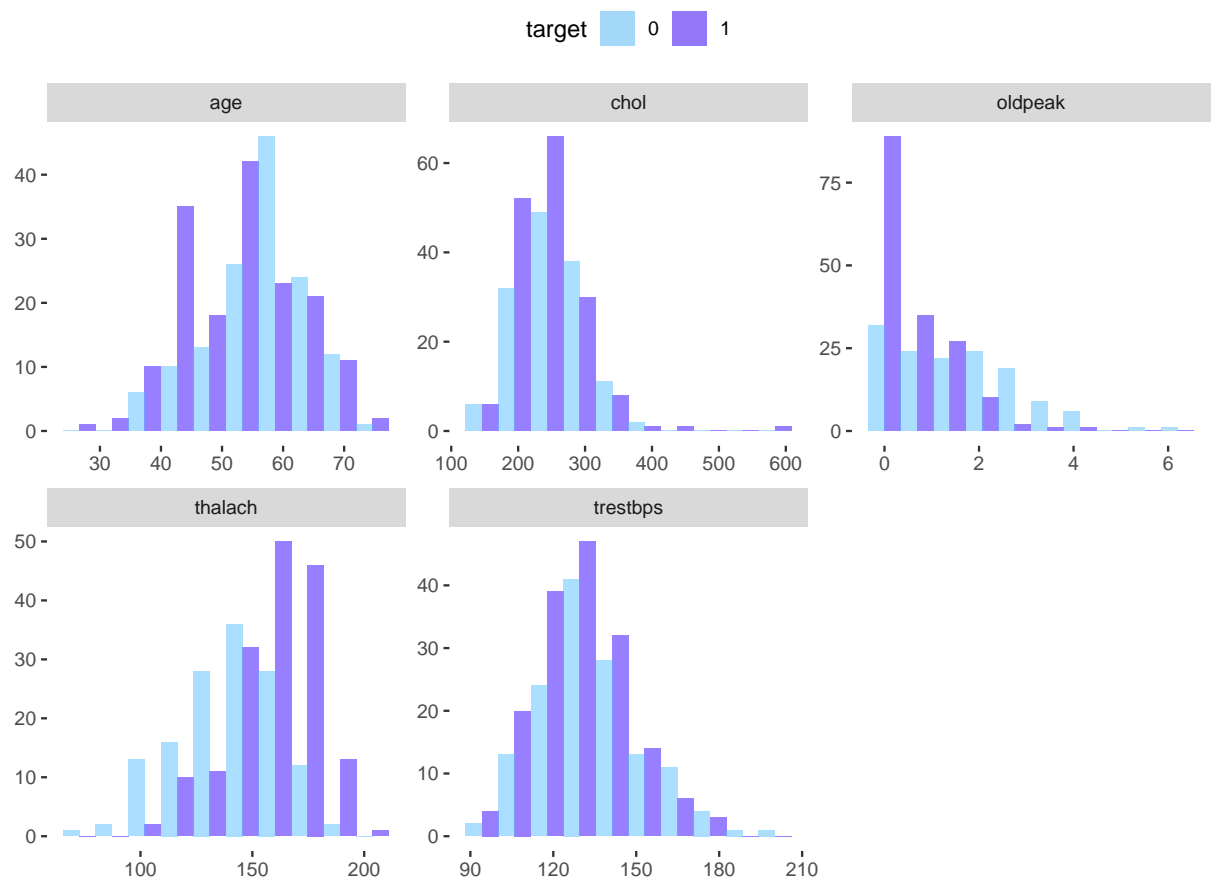


Figure 1: Numeric Data Distributions as a Function of TARGET



Figure 2: Categorical Data Distributions as a Function of TARGET

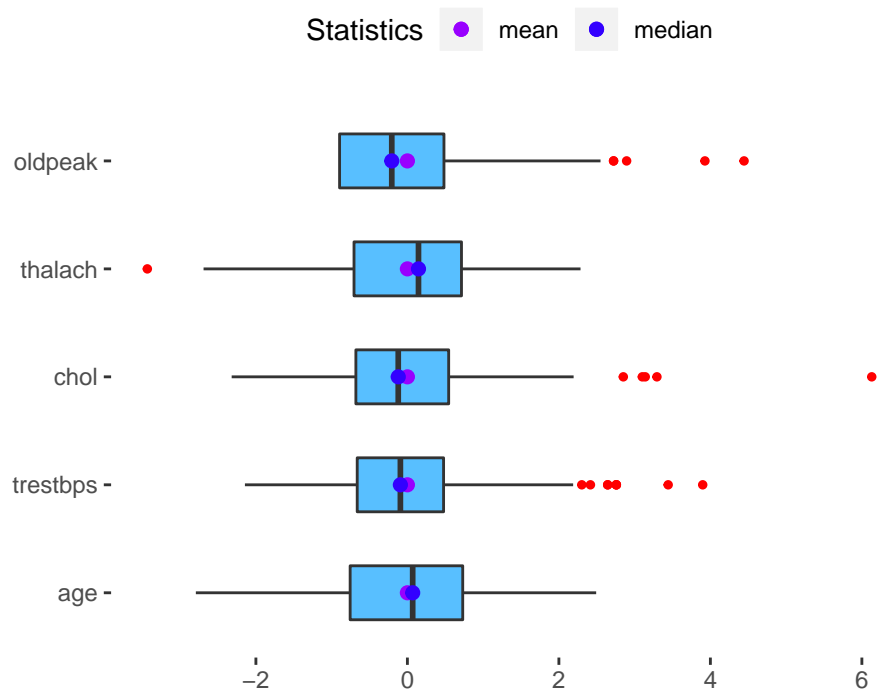


Figure 3: Scaled Boxplots

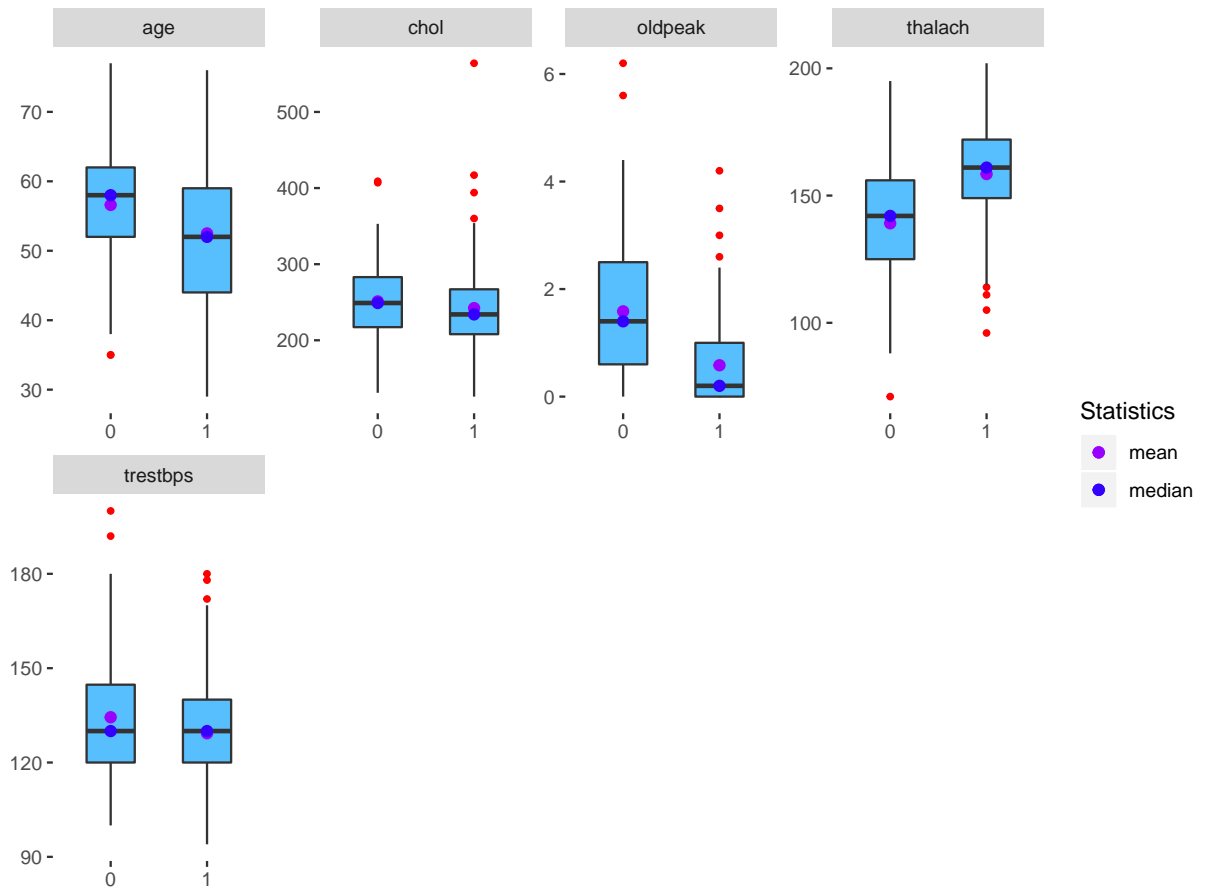


Figure 4: Linear relationship between each numeric predictor and the target

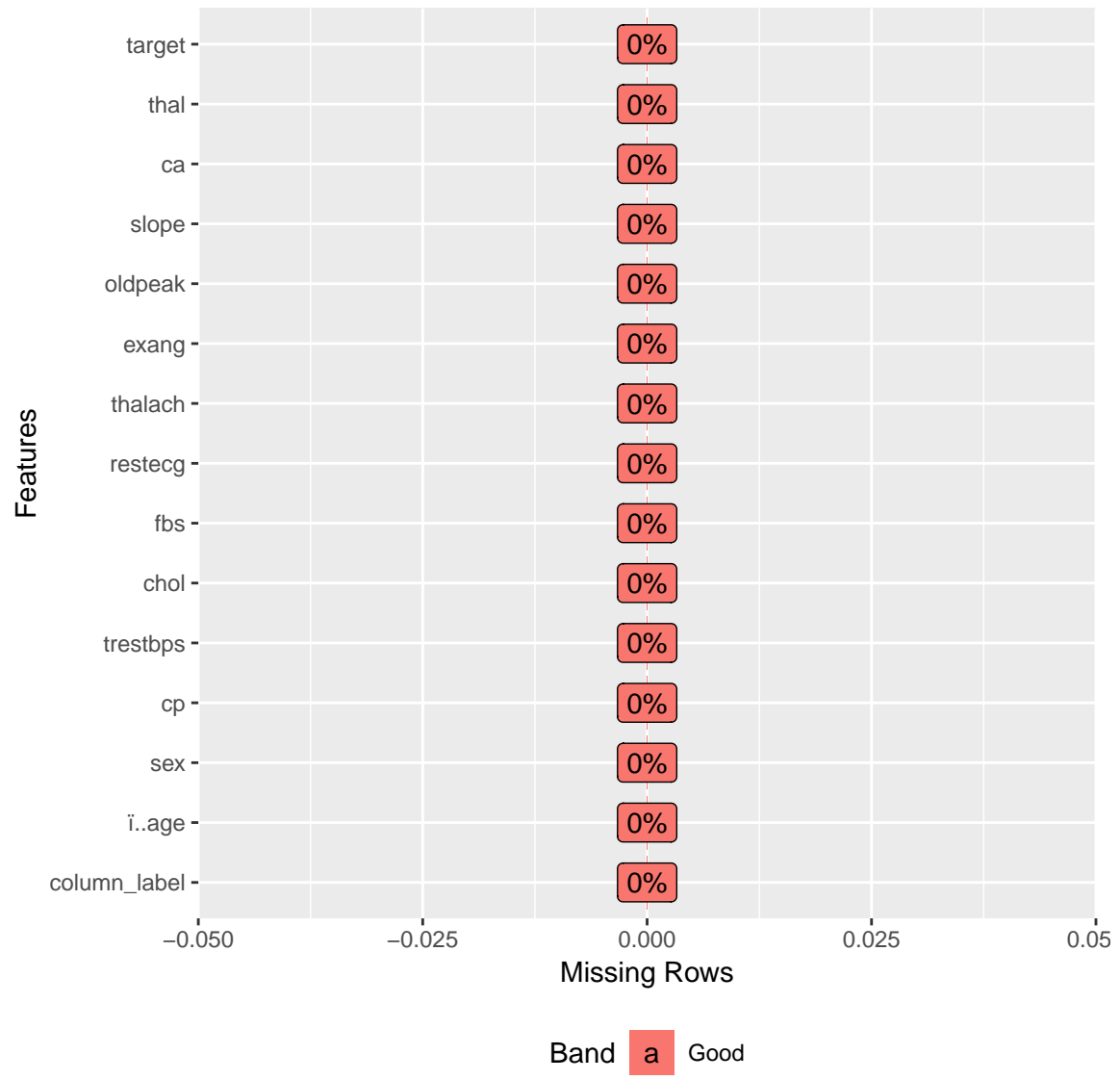
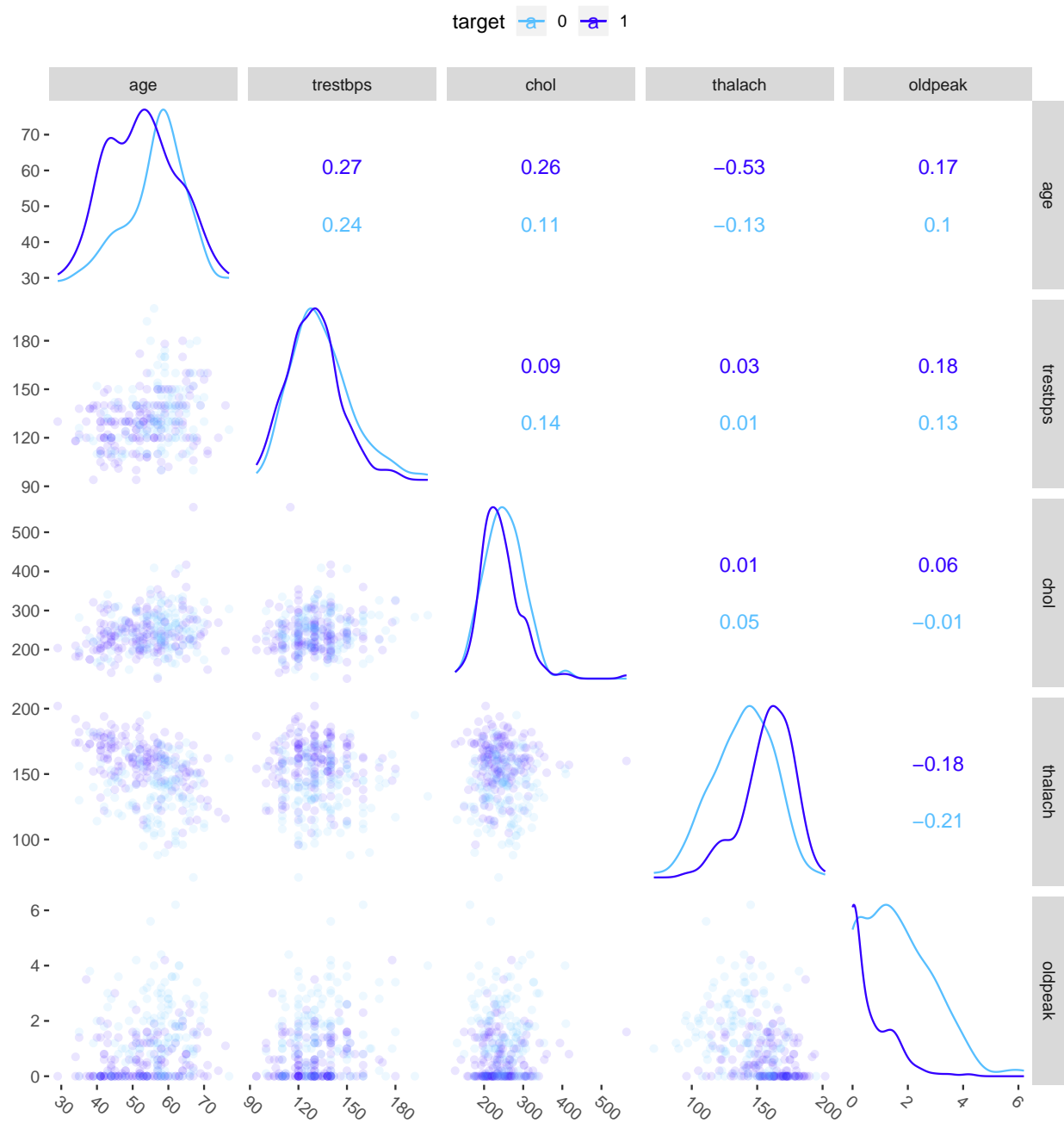


Figure 5: Missing data in the synthesised dataset



6.2 R statistical programming code

The appendix is available as script.R file in `projectFinal_heart` folder.

https://github.com/betsyrosalen/DATA_621_Business_Analyt_and_Data_Mining